Submitted to Eurospeech'99, Budapest

SPEECH/MUSIC DISCRIMINATION BASED ON POSTERIOR PROBABILITY FEATURES

Gethin Williams * *and Daniel P.W. Ellis* † [1]

* Department of Computer Science, Sheffield University, U.K.
† International Computer Science Institute, 1947 Center St, Berkeley, CA 94704
Email: g.williams@dcs.shef.ac.uk, dpwe@icsi.berkeley.edu

ABSTRACT

A hybrid connectionist-HMM speech recognizer uses a neural network acoustic classifier. This network estimates the posterior probability that the acoustic feature vectors at the current time step should be labelled as each of around 50 phone classes. We sought to exploit informal observations of the distinctions in this posterior domain between nonspeech audio and speech segments well-modeled by the network. We describe four statistics that successfully capture these differences, and which can be combined to make a reliable speech/nonspeech categorization that is closely related to the likely performance of the speech recognizer. We test these features on a database of speech/music examples, and our results match the previously-reported classification error, based on a variety of special-purpose features, of 1.4% for 2.5 second segments. We also show that recognizing segments ordered according to their resemblance to clean speech can result in an error rate close to the ideal minimum over all such subsetting strategies.

1. INTRODUCTION

In the hybrid connectionist-HMM framework for automatic speech recognition [1], typical practice is to train a neural network to estimate the posterior probabilities of around 50 context-independent phone classes given a temporal window of approximately 100 ms of feature vectors i.e. $\{p(q_k|X)\}$ where $\{q_k\}$ are the phone labels and X represents the acoustic features. This vector of time-varying probabilities forms a compact description of the analysis by a given acoustic model, and we often precalculate this data. Visualization of these values, as shown in figure 1, has led to informal observations concerning the gross distinctions between segments consisting of clean speech and other segments, such as music or very noisy speech, which are unlikely to be recognized successfully. The work described below seeks to formalize and exploit these observations in order to distinguish speech segments from intervals of nonspeech signal.

In our standard speech recognizer, phonetic classification and subsequent recognition is attempted for all input frames. If the acoustic signal input to the recognizer contains periods of badly corrupted speech or non-speech, this strategy will at times squander valuable computational resources attempting to decode word sequences where none are present. It makes good sense therefore, both in decoding expense and word-error rate (WER) terms, to attempt to excise 'unrecognizable' portions from the incoming acoustic signal.

A popular approach to speech/music discrimination has been to take the decorrelated feature frames used for phone recognition, but to train new distribution models to distinguish training data labelled either as speech or nonspeech (typically music) [2, 3]. One criticism of this approach is that the smoothed spectral surface underlying MFCCs and similar features has been specifically selected to hide and remove aspects of the signal that are not phonetically relevant, such as speaker identity and background noise, and we might expect these features to form a poor basis for distinguishing speech from nonspeech in comparison to specially-devised features [4]. Here, however, we start with the posterior probabilities from the acoustic classifier – features even more specific to speech and further removed from the original signal – as an indicator of the presence or absence of recognizable speech, i.e., to answer the question, "Are there or are there not phones present at this point in the signal?"

One perspective on this apparent contradiction is to think about the acoustic phone classifier as containing a set of highly-tuned detectors for uniquely speech-like signal events. The use of 'hard-targets', where the training targets switch in a single step from being one label to another, encourages the network to 'turn up the gain' around transitions between phone segments, i.e. to sharpen the response in these regions as much as possible, given the information from the acoustic context. As such, the network is encouraged to learn the complex temporal structure of telltale speech gestures present in phones and their transitions. A signal containing a substantial and balanced collection of these gestures is effectively being recognized as speech, and a signal that rarely passes through these critical patches of feature-space fails to show any resemblance to speech. The phone classifier network thus provides a very specific and significant amplification of the distinction between speech and nonspeech segments. Since we have such a tuned "speech-event detector" to hand (and since the input signal is passed through it in any case as the first stage of speech recognition), it is attractive to take advantage of the distinct characteristics revealed.

[1] Joint first authors.

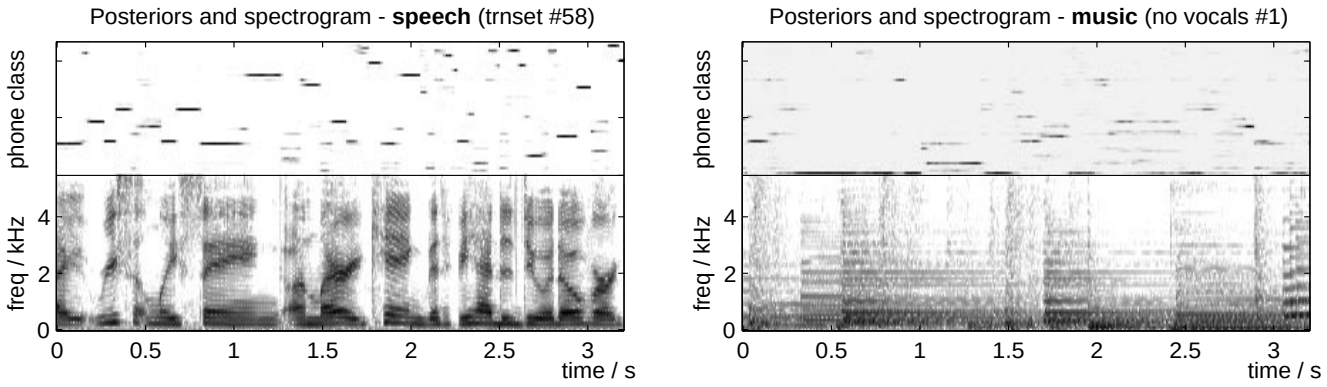

Figure 1: Comparison of an audio segment consisting of clean speech (left) and a segment of instrumental music (right). Each signal is represented by its spectrogram (lower half) and the 54 per-time-frame phone posterior probabilities displayed as rows of an image (upper half).

2. FEATURE DESIGN

We have experimented with four features, as described below. Each one condenses the posterior phone probability array (indexed by phone label and time step) into a single decision variable for each segment:

- Mean per-frame entropy, defined as:

$$\frac{1}{N}\sum_{n=0}^{N-1}\sum_{k} -p(q_k^n).log(p(q_k^n)) \qquad (1)$$

  where n is the frame number (time index), N is the number of frames in the segment, and $p(q_k^n)$ is the posterior probability of label q_k at time n as estimated by the acoustic model.

  Since the phone labels are mutually exclusive, the posterior probabilities at a given time represent a true pdf, and the entropy of that pdf (the expected value of the log probability) is a measure of the goodness-of-fit of the current observations to the acoustic model: A distribution dominated by a large probability for a single label will have an entropy close to zero (meaning little information is gained by knowing its actual value, or, equivalently, there is little uncertainty over the unknown value), whereas the situation in which a number of labels are given roughly equal probabilities (each relatively small, since they must sum to one) will have a much larger entropy.

  We have used per-frame entropy previously for identifying individual words that are poorly modeled [5] and as a basis for choosing segmentation points [6].

- Average probability 'dynamism', defined as:

$$\frac{1}{N}\sum_{n}\sum_{k}(p(q_k^n) - p(q_k^{n-1}))^2 \qquad (2)$$

  This measure is based on the observation that the probability estimates for well-modeled speech segments change both abruptly and frequently, as normal speech moves between phones every few tens of milliseconds. By contrast, signals that are not speech tend to cross these phone boundaries much less frequently and typically more gradually, both of which serve to reduce this measure of the mean-squared first-order difference in phone probabilities.

- Background-label energy ratio, defined as:

$$\frac{\sum_n p(q_{sil}^n).e^n / \sum_n p(q_{sil}^n)}{\sum_n (1-p(q_{sil}^n)).e^n / \sum_n (1-p(q_{sil}^n))} \qquad (3)$$

  Here, q_{sil} is the specific classifier label associated with inter-speech gaps in the training corpus, and e^n is the energy of the original speech signal in a window around time step n. Since the background label is usually trained over a very wide range of signals, it has a tendency to be a catch-all for nonspeech intervals, and is frequently active in signals that are not speech (as well as in speech signals with silent gaps). In a segment of clean, well-recorded speech, segments labelled as background should have much less energy than the other segments (in which speech has been identified), and thus this ratio of the expected energy of background-labelled segments to the expected energy of speech-labelled segments should be very small. By contrast, segments dominated by nonspeech, as well as very noisy speech segments, will typically label high-energy frames as background, bringing the ratio to one or larger.

- Phone distribution match, for instance

$$(S_r - S_{speech})^T \Lambda_{speech}^{-1} (S_r - S_{speech}) \qquad (4)$$

  Where S_r is a vector of statistics for the entire array of probability values over the segment r, in this case the standard deviation along time of each label's probability, weighted to discount background/silence states, e.g.

$$S_r(k) = \sqrt{\frac{\sum_n (p(q_k^n) - \bar{p(q_k)})^2 (1-p(q_{sil}^n))}{\sum_n (1-p(q_{sil}^n))}} \qquad (5)$$

  where $\bar{p(q_k)}$ is the mean probability for label q_k over the segment. These values are assembled into $S_r = \{S_r(k)\}$ over the whole set of labels excluding background/silence. S_{speech} is the expected value of this vector based on a training set of known clean-speech samples, and Λ_{speech} is the covariance matrix of those training vectors, constrained to be diagonal to avoid overfitting (in this case, it is a 54×54 matrix trained on just 60 examples for the 15 second segment examples below).

  To the extent that observed speech segments are long enough to be phoneticaly balanced, it should be possible to capture these underlying patterns in S_{speech}, and

Feature	Sp. acc.	Mus acc.	Error	d'
Entropy	75/80	73/80	7.5%	3.3
Dynamism	80/80	80/80	0%	4.9
Energy	78/80	79/80	1.9%	6.0
Distribution	78/80	80/80	1.3%	4.3
4 features	80/80	80/80	0%	9.6
3 features	80/80	80/80	0%	7.9

Table 1: Classification accuracy for the four different posterior-based features, and for the combined Gaussian model with and without the 'Distribution' feature. The first column is how many of the 15 second speech segments were classified as speech, out of a total of 80; the second column counts music segments classified as nonspeech. The third column expresses these as an overall classification error, and the final column gives the d' measure of class separation.

Feature	Sp. acc.	Mus acc.	Error	d'
Entropy	425/480	402/480	13.9%	1.9
Dynamism	447/480	462/480	5.3%	3.0
Energy	434/480	458/480	7.1%	2.9
Distribution	151/480	444/480	38.2%	0.5
4 features	472/480	472/480	1.7%	4.7
3 features	476/480	472/480	1.3%	4.7

Table 2: As the previous table, but for 2.5 second segments i.e. with each of the 160 Scheirer/Slaney examples divided into 6 equal pieces. No segments were split across training/test sets. Note that excluding the 'Distribution' feature in the final line actually helps performance.

to detect the arbitrarily-different behavior of nonspeech segments. In fact, the specific motivation for this measure was an observation that music segments may sometimes be classified predominantly as a single phone label (such as /n/); a distribution measure such as the one described should highlight such wildly skewed instances.

Each of these metrics has a reasonable ability to discriminate between speech and music segments, as shown in the first four rows of table 1. This reports classification results on a set of 80 speech and 80 music examples, of 15 seconds each, recorded at random from the radio during the summer of 1996 by Scheirer and Slaney [4]. Classification is performed by calculating means and variances separately for the speech and music training examples, then performing a Gaussian likelihood ratio test. The classification results are reported for all 160 segments, jacknifed into four cuts, with three-quarters used to set the parameters and one-quarter used for test in each cut. The posterior probabilities are derived from a combination of three neural-net classifiers: a 256 hidden-unit (HU) recurrent net based on PLP features, an 8000 HU multi-layer perceptron (MLP) trained on modulation-filtered spectrogram features, and an 8000 HU MLP based on PLP features. These models are derived from our recent entry in the DARPA/NIST "Broadcast News" evaluation, and have been trained on approximately 140 hours of speech provided in that task [7]. As the table shows, all measures perform well; the number of errors is so small that relative judgments are hard to make, and the performance differences are barely significant at best. Reporting the d' value for each classifier, that is the distance between the two class means divided by the average within-class standard deviation, gives a less quantized measure of classifier success, but it is not directly related to the classification error owing to the non-Gaussian nature of the data.

Table 2 shows the same results for 2.5 second segments obtained by dividing each segment into six equal pieces and treating each as a separate example. These results are intended to be directly comparable with the 1.4% classification error quoted by Scheirer & Slaney. However, our best result of 1.3% is not significantly different from their result.

The fifth and sixth lines of both tables are obtained by combining metrics (either all four, or the three left after excluding the 'Distribution' measure) into a single indicator. This is done by fitting two 3- or 4-dimensional full-covariance Gaussians to, respectively, the speech and nonspeech training examples in each cut, then classifying the remainder based again on a likelihood ratio test. This combination manages to improve on any of the metrics alone, which is to be expected since they each focus on rather different attributes of the underlying data, and should therefore combine advantageously. In the more challenging case of the shorter segments, however, the 'Distribution' metric performs poorly and actually hurts when added to the combined metric. Presumably, 2.5 seconds is too short a sample to achieve the phonetic balance detected by this measure, whereas 15 seconds is adequate for successful discrimination.

3. APPLICATIONS

How can this ability to discriminate speech and nonspeech segments be used to help the transcription of broadcast audio? When a particular recording contains both speech and music, it is preferable to avoid attempting to recognize segments of nonspeech as speech. Typically, a recognizer will generate a 'nonsense' hypothesis of the word sequence that fitted the data least poorly, and these illusory words are a confusing distraction when interleaved with the 'proper' transcription. Moreover, because of the highly equivocal match, the hidden Markov model decoder will typical expend considerable computational effort on this worthless task, as shown in [6]. Since decoding time typically dominates speech recognition, it is desirable to avoid decoding the nonspeech segments, if they can be quickly identified as such. The measures we have described are all quickly calculated prior to decoding, and so serve this role. To emphasize the point, figure 2 shows the variation in overall word error rate as a function of how many segments are decoded, where the segments that look most like speech and least like music are decoded first. This is over the entire 246 segment set, including speech, music and mixtures of the two; when a pure music segment is decoded, all the resulting words count as errors, since there is no speech in the segments, and hence the 'correct' transcription would have no words. Thus the error rate reaches a minimum roughly half way through the graph, when essentially all the speech segments have been decoded but none of the music segments have been attempted. From then onwards, each new decode simply adds insertion errors, so the rate climbs. In practice, a threshold would be established, aiming to find this minimum overall transcription error; segments falling below this level would be deemed nonspeech and not transcribed.

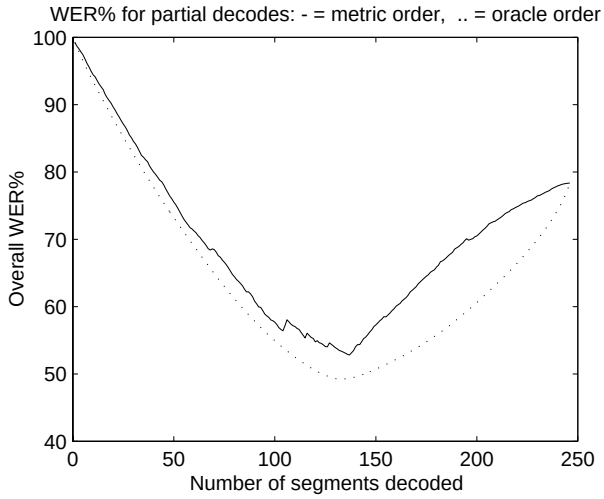

Figure 2: Plot of the variation in overall word error rate with the number of segments decoded, starting with those segments judged most likely to contain clean speech. This is for the entire data set which is composed of the 80 speech and 80 music examples used above plus additional samples including 60 of speech-over-music which include many 'valid' words to be recognized. About half the segments do not contain any words, and decoding them simply worsens the error rate by introducing 'insertion' errors, which is why the curve rises on the right. The dotted line shows the oracle best performance, with the decoder choosing the best possible next segment at each step.

4. CONCLUSIONS

In summary, we have shown that features derived from the phone-posterior estimates are highly effective in distinguishing segments of clean speech from other segments, such as those containing music. Although this scheme is founded upon the acoustic models of the speech recognizer, which were trained only to make phonetic distinctions, it is able to match the performance of the purpose-designed speech/music features described by Scheirer and Slaney. Moreover, this tight coupling to the speech recognition makes the speech/music discrimination highly relevant to the profitability of attempting to recognize words in the segment, so the classifier is particularly valuable in this role.

We note that this work presumes the availability of audio already broken into consistent segments consisting of speech or nonspeech. While many algorithms exist to perform this segmentation, we are now investigating the use of these statistics in this necessary first stage.

5. ACKNOWLEDGMENTS

This work was funded by the European Union through the SPRACH (20077) and THISL (23495) projects. We are also very grateful to Eric Scheirer, Malcolm Slaney and Interval Research Corporation for making available to us their database of speech/music examples.

6. REFERENCES

[1] N, Morgan, and H, Bourlard, "Continuous Speech Recognition: An Introduction to the Hybrid HMM/Connectionist Approach," *Signal Processing Magazine*, pp 25-42, May 1995.

[2] M. Siegler, U. Jain, B. Raj and R. Stern "Acoustic segmentation, classification and clustering of Broadcast News audio," *Proc. DARPA Speech Recognition Workshop, 1997*

[3] M. S. Spina and V. W. Zue "Automatic transcription of general audio data: Preliminary analyses," *Proc. ICSLP, Philadelphia, 1996*

[4] E. Scheirer and M. Slaney "Construction and evaluation of a robust multifeature speech/music discriminator," *Proc. ICASSP, Munich, 1997*

[5] G. Williams and S. Renals "Confidence Measures Derived from an Acceptor HMM," *Proc. ICSLP, Sydney, 1998*

[6] J. Barker, G. Williams and S. Renals "Acoustic Confidence Measures for Segmenting Broadcast News," *Proc. ICSLP, Sydney, 1998*

[7] G. Cook, J. Christie, D. Ellis, E. Fosler-Lussier, Y. Gotoh, B. Kingsbury, N. Morgan, S. Renals, A. Robinson, and G. Williams "The SPRACH System for the Transcription of Broadcast News," *Proc. DARPA Speech Recognition Workshop, 1999*